# EXTENDING THE SCOPE OF LECTOMETRY I:

# FROM DIALECTS TO GLOBAL VARIETIES

# AND

# EXTENDING THE SCOPE OF LECTOMETRY II:

# NEW METHODS AND FEATURES

**ORGANISER(S)**

**Jocelyne Daems**

**Karlien Franco**

**Laura Rosseel**

**Melanie Röthlisberger**

**QLVL, University of Leuven, Belgium**

**Keywords:** ***dialectometry, stylometry, sociolectometry, language perception, aggregated distance measures***

This panel aims to showcase research in the field of lectometry. In this field, quantitative measures are employed to aggregate over linguistic variables in order to establish the relative similarity (or distance) between different lects. These lects are collections of linguistic features that can vary along any extra-linguistic contextual dimension in the broadest sense possible (Geeraerts, Grondelaers and Bakema 1994: 4). Given the definition above, several fields of linguistic research fall within the scope of lectometry. Specifically, in dialectometry, stylometry, sociolectometry and language perception research, distances between lects are studied along the geographical, discursive, social and subjective axis respectively. In this panel, we aim to highlight the range of research questions that can be addressed against the background of lectometry.

Firstly, the geographical axis of lectometry is developed in dialectometry. In traditional dialectometric research, the relative (dis)similarity between dialects is established by aggregating over a large set of dialectal features (e.g. Goebl 2006, Heeringa 2004, Nerbonne and Kleiweg 2003, Séguy 1971, Szmrecsanyi 2013). Recently, however, the field of dialectometry is witnessing a trend of widening its scope from dialects to sociolects (e.g. Hansen 2012, Wieling, Nerbonne and Baayen 2011).

Secondly, stylometry and register analysis are situated along the discursive axis of lectometry. In stylometric studies, the distribution of linguistic features in texts provides insight into the ways in which authors have individual and thus distinguishable styles (e.g. Grieve 2007, Luyckx and Daelemans 2011). Also related to the discursive axis are studies like Biber (1995), which looks into how text types/genres vary, positioning them, for instance, along functional dimensions such as ‘involvedness’ or ‘narration’.

The third field of study related to lectometry, sociolectometry, considers language variation in relation to traditional factors such as age, gender or region. A prime example of a sociolectometric study is Geeraerts, Grondelaers and Speelman (1999), which examines lexical variables in order to measure the relation between the two main national varieties of Dutch. Expanding on this early work in sociolectometry, Speelman, Grondelaers and Geeraerts (2003) and Ruette et al. (2014) introduce more elaborate quantitative techniques such as cluster analysis and multi-dimensional scaling. Advanced methodological techniques, like Semantic Vector Space models in Ruette, Ehret and Szmrecsanyi (2016), have recently been employed in sociolectometry as well.

The fourth field of study that falls within the scope of lectometry, language perception research, is situated along the subjective axis. So far, lectometry has mainly focused on measuring distances between varieties based on language production data. However, measuring subjective distances on the basis of language perception and attitudes would offer a valuable addition. This avenue is still relatively unexplored compared to the three fields above, but studies like Gooskens and Heeringa (2004) or Van Bezooijen and Heeringa (2006) certainly offer a steppingstone to further developing this aspect of lectometry.

To sum up, lectometry offers an interesting umbrella perspective for the aforementioned fields measuring distances between language varieties along different axes. The aim of this panel is to catalogue the range of different lectometric approaches and the ways in which they can reinforce each other. More specifically, research questions include but are not restricted to the following ones:

1. How can insights from different linguistic fields (e.g. Cognitive Linguistics) inform lectometric research?
2. Do text types in contact situations exhibit the same dimensional patterns as in more traditional settings?
3. How does sociolinguistic variation (in the narrow sense) influence dialectometric results?
4. Which methods and datasets are available that can be used to combine different approaches to language variation (e.g. geographical, stylistic and social variation) into one comprehensive framework?
5. Can social psychological attitude measures recently adopted in linguistic perception research (e.g. Speelman et al. 2013, Pantos and Perkins 2012) provide interesting tools to measure subjective distances between languages/language varieties?

This panel is divided in two parts[1] according to how the contributions expand and innovate current research lines in lectometry. The first and present part brings together papers that expand the scope of lectometry from the more traditional dialects to global varieties. The second part of the panel focuses on lectometric research that introduces new methods and linguistic features into the field. The first part is preceded by a short introduction by the organizers and the panel's keynote speaker, Martijn

[1] As agreed with the local organizers, we are submitting a twofold panel. Both parts are to be scheduled one after the other as to ensure maximal interaction between all researchers involved and to emphasize the fact that both parts make up one themed session.

Wieling (Winner of the 2016 European Young Research Award), who will be talking about “Generalized additive modeling as a useful tool for dialectometry.” Both talks will emphasize how the papers in the two parts of the panel are interconnected and invite discussion and interaction between the various strands of research represented by our participants. Our panel concludes with a discussion slot, led by Dirk Geeraerts and Dirk Speelman, which will bring together ideas put forward in both parts of the panel. In addition, there will be a focus on perception research, the subfield of lectometry underrepresented in this panel, and how we can encourage scholars in this field to enter into dialogue with lectometric work.

**References:**


Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambrigde: Cambrigde University Press.

Geeraerts, D., S. Grondelaers and D. Speelman (1999). *Convergentie en divergentie in de Nederlandse woordenschat: een onderzoek naar kleding- en voetbaltermen*. Amsterdam: P.J. Meertens-Instituut.

Geeraerts, D., S. Grondelaers and P. Bakema (1994). *The Structure of Lexical Variation. Meaning, Naming, and Context*. (Cognitive Linguistics Research 5). Berlin/New York: Mouton de Gruyter.

Goebl, H. (2006). Recent advances in Salzburg dialectometry. *Literary and Linguistic Computing* 21(4), 411–435.

Gooskens, C. and W. Heeringa (2004). Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change* 16(3), 189–207.

Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing* 22(3), 251–270.

Hansen, S. (2012). Dialektalität, Dialektwissen und Hyperdialektalität aus soziolinguistischer Perspektive. In S. Hansen, C. Schwarz, P. Stoeckle and T. Streck (eds.). *Dialectological and Folk Dialectological Concepts of Space. Current Methods and Perspectives in Sociolinguistic Research on Dialect Change* (pp. 48–74). Berlin: de Gruyter.

Heeringa, W. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. PhD thesis, Rijksuniversiteit Groningen.

Luyckx, K. and W. Daelemans (2011). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing* 26(1), 35–55.

Nerbonne, J. and P. Kleiweg (2003). Lexical distance in LAMSAS. *Computers and the Humanities* 37(3), 339–357.

Pantos, A. J. and A. W. Perkins (2012). Measuring implicit and explicit attitudes toward foreign accented speech. *Journal of Language and Social Psychology* 32(1), 3–20.

Ruette, T., K. Ehret, B. Szmrecsanyi (2016). A lectometric analysis of aggregated lexical variation in written Standard English with Semantic Vector Space models. *International Journal of Corpus Linguistics* 21(1), 48–79.

Ruette, T., D. Geeraerts, Y. Peirsman and D. Speelman (2014). Semantic weighting mechanisms in scalable lexical sociolectometry. In B. Szmrecsanyi and B. Wälchli (eds.). *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech* (pp. 205–230). Berlin/New York: de Gruyter.

Séguy, J. (1971). La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35(138), 335–357.

Speelman, D., A. Spruyt, L. Impe and D. Geeraerts (2013). Language attitudes revisited: auditory affective priming. *Journal of Pragmatics* 52, 83–92.

Speelman, D., S. Grondelaers and D. Geeraerts (2003). Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities* 37(3), 317–337.

Szmrecsanyi, B. (2013). *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry*. (Studies in English Language). Cambridge: Cambridge University Press.

Van Bezooijen, R. and W. Heeringa (2006). Intuitions on linguistic distance: geographically or linguistically based? In T. Koole, J. Nortier and B. Tahitu (eds.). *Artikelen van de Vijfde Sociolinguïstische Conferentie* (pp. 77–87). Delft: Eburon.

Wieling, M., J. Nerbonne and R. H. Baayen (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE* 6(9), e23613.

**Introduction by panel organizers**

Jocelyne Daems | University of Leuven

Karlien Franco | University of Leuven

Laura Rosseel | University of Leuven

Melanie Röthlisberger | University of Leuven

(5 minutes)

**Keynote: Generalized additive modeling as a useful tool for dialectometry**

Martijn Wieling | University of Groningen

(30 minutes + 5 minutes for questions)

**EXTENDING THE SCOPE OF LECTOMETRY I: FROM DIALECTS TO GLOBAL VARIETIES**

1. **A Quantitative Approach to Swiss German Dialect Syntax**
   Yves Scherrer | Université de Genève
   Philipp Stoeckle | Universität Zürich
2. **Mapping the structure of dialect/standard repertoires: on the use of sociolectometric methods**
   Anne-Sophie Ghyselen | University of Ghent
3. **Measuring language contact in geographical space**
   Xulio Sousa | Universidade de Santiago de Compostela
4. ***SE* constructions in European Portuguese and Brazilian Portuguese and the clitic loss, maintenance and insertion: a corpus-based sociolectometric and socio-cognitive analysis**
   Augusto Soares da Silva | Catholic University of Portugal, Braga
   Clara Vanderschueren | Ghent University
   Dafne Palú | Catholic University of Portugal, Braga
5. **Investigating geographic and register variation in world Englishes**
   Axel Bohmann | The University of Texas at Austin

(All regular presentation slots of 15 minutes + 5 minutes for questions)

**EXTENDING THE SCOPE OF LECTOMETRY II: NEW METHODS AND FEATURES**

1. **A corpus- and profile-based lectometric analysis of emotion concepts in European Portuguese and Brazilian Portuguese**
   Augusto Soares da Silva | Catholic University of Portugal, Braga
2. **Applied lectometry: Using a multivariate spatial analysis to identify cultural regions**
   Jack Grieve | Aston University
3. **The sociolectometry of Flemish online teenage talk: Social and medium-related variation in the use of expressive markers**
   Lisa Hilte | University of Antwerp
   Reinhild Vandekerckhove | University of Antwerp
   Walter Daelemans | University of Antwerp
4. **Probabilistic lectometry**
   Benedikt Szmrecsanyi | University of Leuven
   Melanie Röthlisberger | University of Leuven
5. **Lectometry and latent variables**
   Koen Plevoets | University of Leuven
6. **Characterizing dialect groups: correlation and informativeness associated with linguistic forms**
   Gotzon Aurrekoetxea | Universidad del País Vasco
   Esteve Clua | Universitat Pompeu Fabra
   Aitor Iglesias | Universidad del País Vasco
   Iker Usobiaga | Universidad del País Vasco
   Miquel Salicrú | Universitat de Barcelona

(All regular presentation slots of 15 minutes + 5 minutes for questions)

**Discussion slot**

Discussants: Dirk Geeraerts & Dirk Speelman | University of Leuven

(20 minutes)

# KEYNOTE: GENERALIZED ADDITIVE MODELING AS A USEFUL TOOL FOR DIALECTOMETRY

**Martijn Wieling**

**University of Groningen**

**Keywords:** ***generalized additive modeling, dialectometry, articulatory data, atlas data***

In this presentation I will introduce and explain a relatively new statistical tool, generalized additive modeling (Wood, 2006), which is excellently suited for quantitatively analyzing dialect data. Generalized additive modeling allows the researcher to model flexible (i.e. non-linear) patterns in large datasets. In this presentation, I will illustrate the use of generalized additive modeling by focusing on two types of dialect data. The first type of analysis focuses on modeling the influence of geography on dialect variation. Rather than the usual dialectometric approach of only focusing on the influence of geography, the generalized additive framework allows the researcher to take into account the complex, non-linear influence of geography, while simultaneously taking into account various sociolinguistic predictors, such as gender of age of the speaker. This approach is illustrated by analyzing a large set of Dutch dialect atlas data (Wieling et al., 2011; Ko et al., 2014). The second type of dialect data covered in this presentation is rather new and involves articulatory data, i.e. the movement of tongue and lips during speech. In this part I will focus on a dialect study (Wieling et al., 2015; submitted) conducted onsite at two schools in the Netherlands, one in the north and one further south. The two schools were located on opposite sides of a strong dialect border. While high school pupils were naming different images in their local dialect, their tongue movement trajectories were measured via three sensors attached to the tongue. In this case, using generalized additive modeling allowed us to analyze the non-linear trajectories of all three sensors over time. Our analysis revealed striking differences between the two dialects with a tongue position which was generally further back for the speakers from the north of the Netherlands. As such, this is the first study which has provided quantitative evidence of differences in articulatory settings at the dialect level.

**References:**

Ko, V., M. Wieling, E. Wit, J. Nerbonne and W. Krijnen (2014). Social, geographical, and lexical influence on Dutch dialect pronunciations. *Computational Linguistics in the Netherlands Journal* 4, 29-38.

Wieling, M., J. Nerbonne and R. H. Baayen (2011). Quantitative Social Dialectology: Explaining Linguistic Variation Geographically and Socially. *PLOS ONE* 6(9), e23613.

Wieling, M., F. Tomaschek, D. Arnold, M. Tiede and R. H. Baayen (2015). Investigating dialectal differences using articulography. *Proceedings of ICPhS 2015*, Glasgow, August 10-14.

Wieling, M., F. Tomaschek, D. Arnold, M. Tiede, F. Bröker, S. Thiele, S. N. Wood and R. H. Baayen. Investigating dialectal differences using articulography. Revised version submitted (July 29, 2015) to *Journal of Phonetics*.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton: Chapman & Hall/CRC.

## A QUANTITATIVE APPROACH TO SWISS GERMAN DIALECT SYNTAX

**Yves Scherrer**

**Université de Genève**

**Philipp Stoeckle**

**Universität Zürich**

**Keywords:** ***dialect syntax, socio-demographic variation, Swiss German, dialectometry***

In the last decades, dialectometry has emerged as a new field of dialectology. As this kind of research requires large amounts of data, many dialectometric studies used data from "traditional" dialect atlases (e. g. ALF, AIS, RND) which were collected by investigating representatives of the oldest dialects available in the survey locations (i.e. the so-called NORMs, cf. Chambers and Trudgill 2004: 29). Moreover, these data contained mostly lexical and phonological (and sometimes morphological) variables, while syntactic phenomena are largely absent in traditional atlases.

In this paper we would like to present results of a dialectometric study that focuses on three aspects which have not been given much attention in previous research. The first aspect concerns the research area, German-speaking Switzerland. Although it is one of the liveliest and at the same time best researched dialect areas in Central Europe, until recently (cf. Goebl, Scherrer and Smečka 2013, Scherrer and Stoeckle accepted) there have been very few dialectometric studies in this area (cf. Kelle 2001). The second aspect regards the investigated linguistic level: our analyses are based on syntax data from the *Syntactic Atlas of German-speaking Switzerland* ('Syntaktischer Atlas der deutschen Schweiz', SADS; cf. Glaser and Bart 2015) which were collected between 2000 and 2002 in 383 locations German-speaking Switzerland. A special characteristic of this atlas – which leads to the third aspect we will focus on – lies in the large number of informants and their varying socio-demographic backgrounds. Whereas in traditional atlas projects, generally one or two representatives were interviewed at each survey location, in the SADS a total of almost 3200 informants participated in the survey (i. e. on average about 8 speakers per location). This gives us not only the possibility to work with frequency instead of binary data for each location, but more importantly, this setting allows us to include socio-demographic variables into our analyses.

In other geographic and sociolinguistic contexts, extralinguistic variables other than geography turned out to be important explanatory factors for dialect variation (cf.

Hansen-Morath 2016, Hansen-Morath and Stoeckle 2014). As for German-speaking Switzerland, various studies focusing on single phenomena from the SADS revealed high correlations between syntactic and socio-demographic variation (cf. Stoeckle accepted, Friedli 2012, Richner-Steiner 2011). However, it is still unclear whether this correlation can be observed for aggregated data and what role socio-demographic variables play in explaining syntactic variation.

In order to answer these questions, we will pursue a twofold approach. On the one hand, we will create different subsets with respect to socio-demographic variables and perform dialectometric analyses for each of these subsets. A comparison of the results will help to answer the question whether a change in the geographic dialect structuring can be observed. On the other hand, we will perform regression analyses in order to determine the importance of different extralinguistic factors in explaining linguistic variation. Finally, the results will have to be interpreted in the light of the specific Swiss-German diaglossic situation, where (contrary to many other contexts) change toward both dialectal and standard structures can be observed.

**References:**


ALF: J. Gilléron and E. Édmont (1902–1910). *Atlas linguistique de la France.* Paris: Champion, 9 vol.

AIS: K. Jaberg and J. Jud (1928–1940). *Sprach- und Sachatlas Italiens und der Südschweiz.* Zofingen: Ringier, 8 vol.

RND: E. Blancquaert and W. Pée (1925–1982). *Reeks Nederlandse Dialectatlassen.* Antwerp: De Sikkel, 16 vol.

Chambers, J. K. and P. Trudgill (2004). *Dialectology*. 2nd edition. Cambridge: Cambridge University Press.

Friedli, M. (2012). *Der Komparativanschluss im Schweizerdeutschen: Arealität, Variation und Wandel*. Dissertation Universität Zürich.

Glaser, E. and G. Bart (2015). Dialektsyntax des Schweizerdeutschen. In R. Kehrein, A. Lameli and S. Rabanus (eds.). *Regionale Variation des Deutschen. Projekte und Perspektiven* (pp. 81–107). Berlin/Boston: de Gruyter.

Goebl, Hans, Y. Scherrer and P. Smečka (2013). Kurzbericht über die Dialektometrisierung des Gesamtnetzes des „Sprachatlasses der deutschen Schweiz“ (SDS). In K. Schneider-Wiejowski, B. Kellermeier-Rehbein, J. Haselhuber (eds.). *Vielfalt, Variation und Stellung der deutschen Sprache* (pp. 153–176). Berlin/Boston: de Gruyter.

Hansen-Morath, S. (2016). *Regionale und soziolinguistische Variation im alemannischen Dreiländereck. Quantitative Studien zum Dialektwandel*. Dissertation Albert-Ludwigs-Universität Freiburg.

Hansen-Morath, S. and P. Stoeckle (2014). Regionaldialekte im alemannischen Dreiländereck – ‚objektive‘ und ‚subjektive‘ Perspektiven. In P. Bergmann, K. Birkner, P. Gilles, H. Spiekermann and T. Streck (eds.). *Sprache im Gebrauch: räumlich, zeitlich, interaktional* (pp. 175–192). Heidelberg: Winter.

Kelle, B.. 2001. Zur Typologie der Dialekte in der deutschsprachigen Schweiz: Ein dialektometrischer Versuch. *Dialectologia et Geolinguistica* 9: 9–34.

Richner-Steiner, J. (2011). *'E ganz e liebi Frau'. Zu den Stellungsvarianten des indefiniten Artikels in der adverbiell erweiterten Nominalphrase im Schweizerdeutschen. Eine dialektologische Untersuchung mit quantitativ-geographischem Fokus*. Dissertation Universität Zürich.

Scherrer, Y. and P. Stoeckle (accepted). A quantitative approach to Swiss German – Dialectometric analyses and comparisons of linguistic levels. *Dialectologia et Geolinguistica*.

Stoeckle, P. (accepted). Zur Syntax von *afa* (‚anfangen') im Schweizerdeutschen – Kookkurrenzen, Variation und Wandel. In A. Speyer (ed.), *Syntax aus Saarbrücker Sicht 2. Beiträge der SaRDiS-Tagung zur Dialektsyntax*. Stuttgart: Steiner.

## MAPPING THE STRUCTURE OF DIALECT/STANDARD REPERTOIRES: ON THE USE OF SOCIOLECTOMETRIC METHODS

**Anne-Sophie Ghyselen**

**University of Ghent**

**Keywords:** ***sociolectometry, Dutch, dialect/standard repertoire***

In his by now famous 2005-article, Auer distinguishes five types of dialect/standard constellations in Europe: (1) exoglossic diglossia, (2) medial diglossia, (3) spoken diglossia, (4) diaglossia, and (5) dialect loss repertoires. His theoretical framework has served as a starting point for several European linguists characterising the language repertoires in their research areas (see e.g. Rys and Taeldeman 2007, Gooskens and Kürschner 2009, Hernández-Campoy and Villena-Ponsoda 2009), and has raised interest in generalizable patterns and dynamics. In this paper, I will discuss how (socio)lectometric research can play a pivotal role in attempts to empirically map the range and internal structure of language repertoires on both the level of the individual as on the level of the speech community. Corpus data on the language behaviour of 30 Flemish women in 5 communicative speech contexts (Ghyselen 2016) will serve as input. On the basis of these data, I will illustrate how a multivariate analysis of 31 linguistic variables yields insight in the internal structure, i.e. the components and the distance between those components, of a dialect/standard continuum. Three statistical methods will be reviewed: (1) correspondence analysis, (2) cluster analysis, and (3) multidimensional scaling. It will be shown how these methods are ideally combined and complemented to gain an in-depth understanding of the structure and dynamics of speech repertoires.

**References:**

Auer, P. (2005). Europe's sociolinguistic unity, or: A typology of European dialect/standard constellations. In N. Delbecque, J. van der Auwera and D. Geeraerts (eds.). *Perspectives on variation: Sociolinguistic, Historical, Comparative* (pp. 7–42). Berlin/New York: Mouton De Gruyter.

Ghyselen, A.-S. (2016). *Verticale structuur en dynamiek van het gesproken Nederlands in Vlaanderen: een empirische studie in Ieper, Gent en Antwerpen*. Gent: Universiteit Gent doctoraatsverhandeling.

Gooskens, C. and S. Kürschner (2009). Cross border intelligibility - on the intelligibility of Low German among speakers of Danish and Dutch. *Zeitschrift für Dialektologie und Linguistic* 138, 273–297.

Hernández-Campoy, J. Manuel and J. Andrés Villena-Ponsoda (2009). Standardness and nonstandardness in Spain: dialect attrition and revitalization of regional dialects of Spanish. *International Journal of the Sociology of Language* 196/197, 181–214.

Rys, K. and J. Taeldeman (2007). Fonologische ingrediënten van Vlaamse tussentaal. In D. Sandra, R. Rymenans, P. Cuvelier and P. Van Petegem (eds.). *Tussen taal, spelling en onderwijs. Essays bij het emeritaat van Frans Daems* (pp. 1–8). Gent: Academia Press.

## MEASURING LANGUAGE CONTACT IN GEOGRAPHICAL SPACE

**Xulio Sousa**

**Universidade de Santiago de Compostela**

**Keywords:** ***language contact, aggregate analysis, dialectometry, geolingusitics***

The quantitative analysis of linguistic data has been employed in variationist studies in order to discover relationships between varieties and patterns of behaviour in features that were hidden to traditional methodologies (Goebl 2006). Dialectrometric studies are helping to understand in a more complete manner the spatial organisation of the varieties, similitudes and distances that occur between readings. In the field of variationist studies, this quantitative methodology tends to be applied in order to analyse varieties within a linguistic domain, independently of its extension (Wieling 2011).

Traditionally, dialectology has been concerned with lexical transfers between varieties associated with a single language, with special attention given to the regional and diachronic spread of particular forms. Less often, the discipline examines lexical transfers between varieties attributed to different languages and the spread of new forms over a linguistic area (Haspelmath 2009). This contribution seeks to demonstrate in what manner the dialectrometric procedures can also be employed in order to analyse the contact between linguistic varieties. The procedures popularized by the Salzburg dialectometric school can be employed to detect patterns of spatial distribution for linguistic forms that belong to different varieties (Goebl 2010). The

aggregate analysis of these linguistic variables proves to be especially useful for a more complete description of the linguistic changes produced by contact. The objective of this contribution is to ascertain as to whether it is possible to discover the existence of geographical patterns in the borrowing process (Speelman, Grondelaers and Geeraerts 2003, Thun 2010, Tadmor, Hapselmath and Taylor 2010).

The different opportunities for the employment of quantitative methodologies will be shown with examples taken from geolinguistic research on the Galician linguistic domain from different periods. The demonstration will focus on geolinguistic contact between varieties in the following aspects:

i. Identification of more permeable areas (prone to change)
ii. Identification of more resistant areas (less prone to change)
iii. Identification of non-linguistic variables that influence change.

**References:**


Goebl, H. (2006). Recent advances in Salzburg dialectometry. *Literary and Linguistic Computing* 21(4), 411–435.

Goebl, H. (2010). Dialectometry: Theoretical prerequisites, practical problems, and concrete applications (mainly with examples drawn from the "Atlas Linguistique de la France", 1902-1910. *Dialectologia*. Special Issue I, 63-77.

Haspelmath, M. (2009). Lexical borrowing: concepts and issues. In M. Haspelmath and U. Tadmor (eds.). *Loanwords in the World's Languages: A Comparative Handbook* (pp. 35–54). Berlin: Mouton de Gruyter.

Speelman, D., S. Grondelaers and D. Geeraerts (2003). Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities* 37(3), 317–337.

Tadmor, U., M. Haspelmath and B. Taylor (2010) Borrowability and the notion of basic vocabulary. *Diachronica* 27(2), 226–246.

Thun, H. (2010). Variety complexes in contact: A study on Uruguayan and Brazilian Fronterizo. In P. Auer and J. E. Schmidt (eds.). *Language and Space: An International Handbook of Linguistic Variation* (pp. 706–723). Berlin: Walter de Gruyter.

Wieling, M., J. Nerbonne and R. H. Baayen (2011). Quantitative Social Dialectology: Explaining Linguistic Variation Geographically and Socially, *PLoS ONE* 6(9). e23613.


## *SE* CONSTRUCTIONS IN EUROPEAN PORTUGUESE AND BRAZILIAN PORTUGUESE AND THE CLITIC LOSS, MAINTENANCE AND INSERTION: A CORPUS-BASED SOCIOLECTOMETRIC AND SOCIOCOGNITIVE ANALYSIS


**Augusto Soares da Silva**

**Catholic University of Portugal, Braga**

**Clara Vanderschueren**

**Ghent University**

**Dafne Palú**

**Catholic University of Portugal, Braga**

**Keywords:** ***constructional variation, impersonal/passive se constructions, clitics, sociolectometry, European and Brazilian Portuguese***

European Portuguese (EP) and Brazilian Portuguese (BP) significantly differ in the use of clitic *se* constructions. EP frequently makes use of a passive *se* construction with agreement (1) and an impersonal *se* construction without agreement (2). In contrast, the general trend in BP is to avoid the clitic *se*: either the accusative *se* of the passive construction, or the nominative *se* of the impersonal construction, are suppressed in cases like (3). The same happens in other uses of the clitic *se* (reflexives, anticausatives and middles). The *se* pronoun deletion, more pronounced in the informal register, has been attributed to the on-going loss of clitics in BP. As an alternative, BP also uses overt personal subject pronouns (*você*, *a gente*, *nós*) instead of the impersonal *se* construction in these contexts (4), which is attributed to the on-going loss of the null subject in BP (Duarte 1995, Kato 1999, Barbosa et al. 2001). A third alternative construction in BP, is the (less frequent) *se* construction without agreement in (5), which is ambiguous between the passive reflexive (1) and the impersonal (2) construction (Duarte et al. 2001).

(1) *Vendem-se casas.*
sell.PRES.3pl-SE houses
(2) *Vende-se casas.*
sell.PRES.3sg-SE houses
(3) *Vende casa(s).*
sell.PRES.3sg house(s)
(4) *A gente vende casa(s).*
people sell.PRES.3sg house(s)
(5) *Se vende casa(s).*
SE sell.PRES.3sg house(s)
'Houses are sold'

Conversely, formal BP tends to insert the clitic *se* in impersonal infinitival constructions, where EP tends towards non-realization (6). In these contexts, the clitic insertion is a strategy to explicitly indicate subject indetermination, which has equally been linked to the increase of the overt subjects in BP (Galves 1987, Colsato 2007).

(6) *É impossível se/Ø achar lugar aqui.* (BP/EP)
be.PRES.3sg impossible SE/Ø find.INF place here
'It's impossible to find a place here'

Based on a corpus of Portuguese and Brazilian texts of the 1950s, 1970s and 2000s, pertaining to different registers (newspapers and magazines, football chats and

blogs), we propose a sociolectometric analysis of the *se* constructions in order to measure the relative (dis)similarity between the two national varieties along the geographical, social, stylistic and historical axes, as well as a socio-cognitive analysis of the conceptual, structural and social factors determining the variation of *se* constructions in EP and BP. The present case study on constructional lectal variation follows the Cognitive Linguistics framework, specifically Cognitive Sociolinguistics (Kristiansen and Dirven 2008; Geeraerts et al. 2010) and is an extension of our sociolectometric and sociocognitive studies on lexical convergence and divergence between EP and BP (Soares da Silva 2010, 2014). Firstly, we shall identify the distributional contexts and meanings of the *se* constructions and, in a similar fashion, of the loss, maintenance and insertion of the clitic se. We shall then analyse the semasiological, onomasiological and lectal variation of the se constructions, developing a usage-feature analysis in order to identify the conceptual, structural and lectal factors of such constructional variation. Conceptually, *se* constructions profile the change-of-state undergone by the thematic participant, and therefore the initiating force is present only in highly schematic terms (Maldonado 2007). The different *se* constructions constitute a continuum of increasing focal prominence of the schematic initiating force profiled as Figure, as in the impersonal construction, or, inversely, of the event terminal point, as in the passive construction. Semasiological and onomasiological profiles of *se* constructions and profile-based sociolectometric measures, i.e. uniformity and featural measures (Geeraerts et al. 1999, Speelman et al. 2003) are used to calculate both the synchronic distance and the diachronic convergence/divergence between EP and BP. Clustering techniques and logistic regression analysis serve to chart the correlation between the conceptual, structural and lectal variables.

**References:**


Barbosa, P., M. Kato and M. E. Duarte (2001). A distribuição do sujeito nulo no português europeu e no português brasileiro. In C. Correia and A. Gonçalves (eds.). *Actas do XVI Encontro Nacional da Associação Portuguesa de Linguística* (pp. 539–550). Lisboa: APL.

Colsato, A. (2007). *A Inserção do SE em Sentenças Não-Finitas do PB*. Dissertação de Mestrado. Universidade de São Paulo.

Duarte, M. E. (1995). *A Perda do Princípio "Evite Pronome" no Português Brasileiro*. Tese de Doutoramento. Campinas: UNICAMP.

Duarte, M. E., M. Kato and P. Barbosa (2001). Sujeitos indeterminados em PE e PB. In *Anais do II Congresso Internacional da ABRALIN* (pp. 405–409).

Geeraerts, D., S. Grondelaers and D. Speelman (1999). *Convergentie en Divergentie in de Nederlandse Woordenschat*. Amsterdam: Meertens Instituut.

Geeraerts, D., G. Kristiansen and Y. Peirsman (eds.) (2010). *Advances in Cognitive Sociolinguistics*. Berlin/New York: De Gruyter.

Galves, C. (1987). *A sintaxe do português brasileiro. Ensaios de Lingüística* 13, 31–50.

Kato, M. (1999). Strong and weak pronominals and the null subject parameter. *Probus* 11, 1–37.

Kristiansen, G. and R. Dirven (eds.) (2008). *Cognitive Sociolinguistics: Language variation, cultural models, social systems*. Berlin/New York: De Gruyter.

Maldonado, R. (2007) Grammatical voice in Cognitive Grammar. In D. Geeraerts and H. Cuyckens (eds.). *The Oxford Handbook of Cognitive Linguistics* (pp. 829–868). Oxford: Oxford University Press.

Nunes, J. (1990). *O Famigerado SE: uma análise sincrônica e diacrônica das construções com se apassivador e indeterminador*. Dissertação de Mestrado. Campinas: UNICAMP.

Soares da Silva, A. (2010). Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese. In D. Geeraerts, G. Kristiansen and Y. Peirsman (eds.). *Advances in Cognitive Sociolinguistics* (pp. 41–83). Berlin/New York: De Gruyter.

Soares da Silva, A. (2014). The pluricentricity of Portuguese: A sociolectometrical approach to divergence between European and Brazilian Portuguese. In Augusto Soares da Silva (ed.), *Pluricentricity: Language variation and sociocognitive dimensions* (pp. 143–188). Berlin/Boston: De Gruyter.

Speelman, D., S. Grondelaers and D. Geeraerts (2003). Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities* 37, 317–337.

## INVESTIGATING GEOGRAPHIC AND REGISTER VARIATION IN WORLD ENGLISHES

**Axel Bohmann**

**The University of Texas at Austin**

**Keywords:** ***World Englishes, register variation, International Corpus of English (ICE), Twitter***

In research on World Englishes, individual national varieties are typically grouped according to their sociolinguistic history (Schneider 2007) and the norm-orientation of the English used in a given country (Kachru 1986). Structural comparison between varieties has mostly been carried out in studies on individual features in a limited number of settings. Insightful as such studies are, it is difficult to make generalizations about the structural relations among World Englishes on their basis.

Feature-aggregation-based approaches to linguistic variation promise to help draw a more systematic picture of unity and diversity in English world-wide. Biber's (1988) multi-dimensional technique has proven instructive in establishing dimensions of variation across registers. More recently, aggregation-based methods have been utilized in the study of regional and typological variation (Szmrecsanyi and Wälchli 2014; Szmrecsanyi 2013; Grieve 2016), but a systematic, empirical application of this

approach to World Englishes has thus far not been attempted (although see Neumann and Fest 2016, Schaub 2016 for steps in this direction).

In this study, I present an aggregation-based study of eight national varieties of English on the basis of 56 morpho-syntactic and lexical features in naturalistic language data. A total of N=6,000 text samples, taken from the International Corpus of English (ICE) and a corpus of geo-located Twitter messages, are coded for their frequency profile for each feature. Factor analysis is then performed on the resulting data matrix to establish the higher-level dimensions structuring the variation in the dataset. Moreover, network diagrams are created to visualize the relationship among the different varieties (cf. Szmrecsanyi 2014: 97-99), based on the aggregate frequencies for all text samples representing each variety, on the whole as well as for individual registers (as reflected in the different ICE text categories).

Results indicate that, while a geographic signal can be traced in the data, the dimensions derived from factor analysis most clearly reflect the communicative properties of different registers, a finding that is in line with Biber (1995). When considering variety differences within individual text categories, the Twitter messages yield the clearest signal. This is most likely due to the fact that these text samples constitute a less coherent register than the ICE samples, and that they are less subject to the homogenizing force of the linguistic standard. The relationship among varieties, as reflected in the different network diagrams, can primarily be understood as a difference between L1 varieties with a long history of codification and more recently emerging L2 varieties.

The study demonstrates that a lectometric approach to World Englishes produces valid results. These may help to put observations from studies of individual features into a broader context of inter-varietal relationships. One question that remains is to what extent it is warranted to discuss differences in varieties on the whole, when these differences are as heavily mediated by register as the present study suggests. Research in World Englishes, whether from a single-feature or an aggregational perspective, will benefit from developing more explicit methods for incorporating register as a factor in its analysis of cross-varietal differences.

**References:**


Biber, D. (1988). *Variation across speech and writing*. Cambridge et al.: Cambridge University Press.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Grieve, J. (2016). *Regional variation in written American English*. Cambridge University Press.

Kachru, B. (1986). *The alchemy of English: The spread, functions, and models of non-native Englishes*. Oxford: Pergamon Press.

Neumann, S. and J. Fest (2016). Cohesive devices across registers and varieties: The role of medium in English. In C. Schubert and C. Sanchez- Stockhammer (eds.). *Variational text linguistics: Revisiting register in English* (pp. 195–220). Berlin/Boston: De Gruyter Mouton.

Schaub, S. (2016). The influence of register on noun phrase complexity in varieties of English. In C. Schubert and C. Sanchez-Stockhammer (eds.). V*ariational text linguistics: Revisiting register in English* (pp. 251–270). Berlin/Boston: De Gruyter Mouton.

Schneider, B. (2007). *Postcolonial English: Varieties around the world*. Cambridge: Cambridge University Press.

Szmrecsanyi, B. (2013). *Grammatical variation in British English dialects: A study in corpus-based dialectometry*. Cambridge: Cambridge University Press.

Szmrecsanyi, B. (2014). Forests, trees, corpora, and dialect grammars. In B. Szmrecsanyi and B. Wälchli (eds.). *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech* (pp. 89–112). Berlin: Mouton De Gruyter.

Szmrecsanyi, B. and B. Wälchli (eds.) (2014). *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech*. Berlin: Mouton De Gruyter.

## A CORPUS- AND PROFILE-BASED LECTOMETRIC ANALYSIS OF EMOTION CONCEPTS IN EUROPEAN PORTUGUESE AND BRAZILIAN PORTUGUESE

**Augusto Soares da Silva**

**Catholic University of Portugal, Braga**

**Keywords:** ***lexical variation, emotions, cultural conceptualization, lectometry, European and Brazilian Portuguese***

In this study we develop a corpus- and profile-based lectometric analysis of three emotion concepts, namely ANGER, PRIDE and LOVE in European Portuguese (EP) and Brazilian Portuguese (BP). The present analysis is part of a wider project on the conceptualization of emotions in EP and BP. The main goal is both to measure the lexical-semantic (dis)similarity regarding emotion concepts between the two national varieties of Portuguese along geographical, social and stylistic axes and to correlate the lectal distances with conceptual and cultural similarities and differences. In order to carry out this lectometric and socio-cognitive study, we follow the Cognitive Linguistics framework, specifically Cognitive Sociolinguistics (Kristiansen and Dirven 2008, Geeraerts et al. 2010) and Quantitative Cognitive Semantics (Glynn and Fischer 2010, Glynn and Robinson 2014), particularly its application to emotion concepts (e.g. Glynn 2007, 2014; Krawczak 2014), and we adopt the sociolectometric methodology developed by Geeraerts et al. (1999), Speelman et al. (2003), Soares da Silva (2010, 2014), Ruette (2012), Ruette et al. (2014). The lectometric analysis uses a concept-, profile-based methodology, where *profile* stands for the set of usage features of a linguistic form or meaning (semasiological profile, also called behavioral profile, Gries 2010) or the set of semantically equivalent usage words in a conceptual category

(onomasiological profile). Profile-based uniformity and featural measures quantify the distance between language varieties. Multivariate statistical techniques, namely multiple correspondence and logistic regression analyses serve to identify emotion usage patterns across the data and to determine their descriptive accuracy and predictive power.

The corpus includes Portuguese and Brazilian texts from blogs and newspapers, compared for stylistic distance measurement. An analysis of a sample of 2500 examples of ANGER (expressed by the lexemes *raiva* 'anger', *fúria* 'fury', *ira* 'anger/wrath', *cólera* 'anger/wrath', *irritação* 'irritation'), PRIDE (lexemes *orgulho* 'pride', *vaidade* 'vanity') and LOVE (lexemes *amor* 'love', *paixão* 'passion', *desejo* 'desire', *atração* 'attraction', *coração* 'heart') will be conducted. The different socio-semantic factors that are associated to the arguments of ANGER, PRIDE and LOVE event-frames, namely Emoter, Cause, Responsible and Receiver will be analyzed. These socio-semantic factors include Emoter behavior and control, Cause type and control, Receiver type, intensity, emotional attitudes, and evaluation (the usage feature analysis is inspired by work in social psychology on emotions, e.g. Fontaine et al. 2013). Different clusters of usage features will be identified. Multiple correspondence analysis shows three clusters of ANGER features: a violent type of anger associated with norm violations and immoral behavior, a complaining type of anger associated with inconveniences, and interpersonal anger associated with the behavior of known people. Two clusters of PRIDE features were found: a self-centered pride and an other-directed pride. Logistic regression reveals some lectal predictors. For instance, belonging to a group or family causes of pride are predictors for EP, whereas the BP predictor is cause relevance for Emoter. This means that EP appears to be more akin to the cluster of other-directed pride, whereas BP seems closer to self-centered pride. As for anger, EP is more consistent with the violent type of anger caused by norm violations and immoral behavior, whereas BP is more associated with the irritating kind of anger caused by inconveniences. These results are in line with cultural conceptualization differences, i.e. the more *collectivist* Portuguese culture in contrast with the more *individualistic* Brazilian culture (Hofstede 2001). In order to measure the lexical-semantic distance between the two national varieties of Portuguese, onomasiological profiles of ANGER, PRIDE and LOVE are also analyzed. In fact, synonyms, mainly denotational synonyms often display sociolinguistic differences and therefore the competition between language varieties.

**References:**


Fontaine, J. R. J., K. R. Scherer and C. Soriano (2013). *Components of Emotional Meaning. A Sourcebook*. Oxford: Oxford University Press.

Geeraerts, D., S. Grondelaers and D. Speelman (1999). *Convergentie en Divergentie in de Nederlandse Woordenschat*. Amsterdam: Meertens Instituut.

Geeraerts, D., G. Kristiansen and Y. Peirsman (eds.) (2010). *Advances in Cognitive Sociolinguistics*. Berlin/New York: De Gruyter.

Glynn, D. (2007). *Mapping Meaning. Toward a Usage-based Cognitive Semantics*. PhD dissertation. Leuven: University of Leuven.

Glynn, D. (2014). The social nature of anger: Multivariate corpus evidence for context effects upon conceptual structure**.** In P. Blumenthal, I. Novakova and D. Siepmann (eds.). *Emotions in Discourse* (pp. 69–82). Frankfurt: Peter Lang.

Glynn, D. and K. Fischer (eds.) (2010). *Quantitative Cognitive Semantics: Corpus-Driven Approaches*. Berlin/New York: De Gruyter.

Glynn, D. and J. Robinson (eds.) (2014). *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*. Amsterdam/Philadelphia: John Benjamins.

Gries, S. Th. (2010). Behavioral Profiles: A fine-grained and quantitative approach in Corpus-based Lexical Semantics. *Mental Lexicon* 5, 323–346.

Hofstede, G. (2001). *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations across Nations*. Thousand Oaks, CA: Sage.

Krawczak, K. (2014). Shame, embarrassment and guilt: Corpus evidence for the cross-cultural structure of social emotions. *Poznan Studies in Contemporary Linguistics* 50(4), 441–475.

Kristiansen, G. and R. Dirven (eds.) (2008). *Cognitive Sociolinguistics: Language variation, cultural models, social systems*. Berlin/New York: De Gruyter.

Ruette, T. (2012). *Aggregating Lexical Variation: Towards large-scale lexical lectometry*. PhD thesis, University of Leuven.

Ruette, T., D. Geeraerts, Y. Peirsman and D. Speelman (2014). Semantic weighting mechanisms in scalable lexical sociolectometry. In B. Szmrecsanyi and B. Wälchli (eds.). *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech* (pp. 205–230). Berlin/New York: De Gruyter.

Soares da Silva, A. (2010). Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese. In D. Geeraerts, G. Kristiansen and Y. Peirsman (eds.). *Advances in Cognitive Sociolinguistics* (pp. 41–83). Berlin/New York: De Gruyter.

Soares da Silva, A. (2014). The pluricentricity of Portuguese: A sociolectometrical approach to divergence between European and Brazilian Portuguese. In A. Soares da Silva (ed.), *Pluricentricity: Language variation and sociocognitive dimensions* (143–188). Berlin/Boston: De Gruyter.

Speelman, D., S. Grondelaers and D. Geeraerts (2003). Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities* 37, 317–337.

## APPLIED LECTOMETRY: USING A MULTIVARIATE SPATIAL ANALYSIS TO IDENTIFY CULTURAL REGIONS

**J. Grieve**

**Aston University**

**Keywords:** ***corpus linguistics, dialectometry, English, lexical variation, social media***

The number and location of American cultural regions has long been the subject of debate. Numerous competing theories have been proposed, which have taken into consideration a long list of different factors, including settlement, ethnicity, religion, and politics. It is difficult, however, to choose between these theories because they have been based almost entirely on the opinion of geographers and historians. Even when empirical data, such as Census records, are taken into consideration, the selection, weighting, and aggregation of these different factors has been subjective. For example, although there can be no doubt that religion is an important factor for defining cultural regions, it is unclear how important this factor is and if its importance is the same across the United States. Assuming, however, that important cultural patterns are reflected in everyday language use, specifically in the topics that people choose to discuss, then the analysis of large regionalized corpora provides an alternative and more objective method for identifying cultural regions.

In this paper, I show how methods borrowed from dialectometry can be used to identify modern American cultural regions. In particular, I analyze the relative frequency of the 10,000 most common words in an 8.9 billon word corpus of geocoded American Tweets collected between 2013 and 2014. To discover common patterns of regional lexical variation in this dataset, I subject the maps for these words to a multivariate spatial analysis, identifying 5 dimensions of lexical variation. I then interpret each of these dimensions regionally, by mapping the dimension scores, and thematically, by classifying the words associated with each dimension by topic. This analysis not only reveals clear regional patterns that align with well-established cultural distinctions, but it also allows for the topics of discussion that characterize language originating from these regions to be identified, including not only topics related to factors traditionally used to identify cultural regions such as religion and ethnicity, but also new factors such as a focus on friendship, family, lifestyle, and the outdoors. Finally, based on these dimensions of lexical variation, I generate a single overall map of American cultural regions, identifying 5 main regions—the Northeast, the Southeast, The Midwest, the South Central, and the West—which both support and challenge previous theories.

In addition to mapping American cultural regions, I also consider what these results tell us about dialect variation. The cultural regions I identify correspond closely to American dialect regions, supporting the theory that dialect regions reflect cultural regions. Although this is not a new theory, the results of this study offer a new explanation for why this relationship holds, as it shows that cultural variation is expressed through differences in the topics that people tend to use language to discuss. This result suggests that regional variation in linguistic structure is not primarily due to arbitrary language change but rather to cultural differences in the way language is used—a hypothesis that challenges basic assumptions underlying sociolinguistic inquiry.

# THE SOCIOLECTOMETRY OF FLEMISH ONLINE TEENAGE TALK: SOCIAL AND MEDIUM-RELATED VARIATION IN THE USE OF EXPRESSIVE MARKERS

**Lisa Hilte**

**University of Antwerp**

**Reinhild Vandekerckhove**

**University of Antwerp**

**Walter Daelemans**

**University of Antwerp**

**Keywords:** ***computer-mediated communication, adolescents, expressiveness, social correlates, computational sociolinguistics***

Expressive markers often function as compensational pragmatic features in informal computer-mediated communication (CMC). The present study analyzes to what extent their use in informal CMC produced by Flemish adolescents correlates with social and medium-related variables, or, in other words, to what extent they are (more or less) prominent markers of 'social digi-lects'.

Our analyses include three types of expressive markers: a number of typographic chatspeak features, an onomatopoeic and a lexical variable. While the research design and the interpretation of the results are essentially sociolinguistic, we rely on computational linguistics methodology for data processing and feature extraction.

The corpus consists of two parts and covers nearly ten years of Flemish adolescent CMC. The first part of the corpus, i.e. the reference corpus for the present study, consists of 2 million tokens and contains chat conversations produced between 2007 and 2013. The social variables that are operationalized are the chatters' gender and age. As for medium, we take synchronicity into account, as well as the public versus private character of the messages. Our general quantitative findings are that girls outperform boys in the expression of emotional involvement (see also Parkins 2012), and younger adolescents outperform the older group. The results are extremely consistent in this respect: the same tendencies can be observed for each of the expressive features. Quite strikingly however, medium has the largest impact: much more expressive markers are used in (largely public) asynchronous social media posts than in (private) synchronous instant messaging. Apart from that, the qualitative analyses lay bare distinct preferences for particular features. E.g. girls prefer other emoticons than their male peers. In other words, expressiveness takes different forms in girls' CMC than in boys' CMC (see Hilte, Vandekerckhove and Daelemans forthcoming).

For the second and more recent part of the corpus, we'll report on a follow-up

study. The new corpus has been collected in 2015 and 2016, which adds a diachronic dimension to the research on expressiveness. While the social variables of age and gender have been maintained, an extra one is added: the educational background and social class of the informants. For educational background, we make a distinction between the three main types of Belgian secondary education, while the social class categorization is based on a cluster of parameters. Medium is no longer a variable in the new data, as all messages are synchronous and private. While we hypothesize that the new analyses will corroborate the quantitative gender and age tendencies of the reference study, we definitely expect qualitative differences, as CMC and youth language are subject to constant renewal, and new technology and media trigger different expressive markers.

Finally, the present research may demonstrate there is no such thing as a Flemish informal adolescent digilect. There are numerous and constantly changing social digilects or digilectal varieties. While most adolescents have access to the very same pool of expressive markers, gender and age determine their preferences, and so does the digital medium and potentially also their educational and social class background.

**References:**

Hilte, L., R. Vandekerckhove and W. Daelemans (2016). Expressiveness in Flemish online teenage talk: A corpus-based analysis of social and medium-related linguistic variation. *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities* (pp. 30–33).

Parkins, R. (2012) Gender and emotional expressiveness: An analysis of prosodic features in emotional expression. *Griffith Working Papers in Pragmatics and Intercultural Communication* 5(1), 46–54.

## PROBABILISTIC LECTOMETRY

**Benedikt Szmrecsanyi**

**University of Leuven**

**Melanie Röthlisberger**

**University of Leuven**

**Keywords:** ***probabilistic constraints, dative alternation, world Englishes, dialectometry, stylometry***

Classical lectometry is based on the joint analysis of variation in feature *frequencies*. Note, however, that the focus on frequencies in classical lectometry is not entirely compatible with the spirit of a neighboring sub-discipline, variationist (socio)linguistics, in which analysts are primarily concerned with the *probabilistic conditioning* of variants,

rather than with their overall text frequency. Against this backdrop, we aim to sketch in this paper a lectometrical methodology that takes as input not frequencies but probabilistic constraints on variant usage. As a case study, we showcase how constrains on syntactic choices structure lectal variation in space and across registers (the analysis is thus an exercise both in dialectometry and in stylometry). Our dataset (see Röthlisberger in preparation) spans >10,000 richly annotated observations of the dative alternation, i.e. variation between the ditransitive dative, as in (1), and the prepositional dative, as in (2), in contexts where both options are possible.

(1) the ditransitive dative variant
That will give [the panel]$_{recipient}$ [a chance]$_{theme}$ to expand on what they've been saying. <ICE-GB:S1B-036>

(2) the prepositional dative variant
[...] and that gives [a chance]$_{theme}$ [to Bhupathy]$_{recipient}$ to equalise the points at thirty all. <ICE-IND:S2A-019>

Observations are drawn from two corpora, the International Corpus of English (Greenbaum 1991), and the Corpus of Web-based English (Davies and Fuchs 2015). We cover in all 9 international varieties of English (British, Canadian, Irish, New Zealand, Jamaican, Singapore, Indian, Hong Kong, and Philippine English), as well as three major registers (spoken language, written language, web language), which yields $9 \times 3 = 27$ lectal datapoints (e.g. spoken British English, web-based Jamaican English). Crucially, rather than basing the lectometrical analysis on the frequencies of particular dative variants in particular sub-corpora, we fit separate regression models per lectal datapoint and aggregate the regression coefficients of well-known constraints on dative choice. For example, it is well-known that animate recipients favor the ditransitive variant (Bresnan et al. 2007). But regression analysis reveals that the effect is stronger in some lects than in others, and it is probabilistic information such as this– along with other measurements such as e.g. the strength of end-weight effects, of pronominality of the recipient, and so on – that feeds into our analysis.

**References:**


Bresnan, J., A. Cueni, T. Nikitina and R. H. Baayen (2007). Predicting the Dative Alternation. In G. Boume, I. Krämer and J. Zwarts (eds.). *Cognitive Foundations of Interpretation* (pp. 69–94). Amsterdam: Royal Netherlands Academy of Science.

Davies, M. and R. Fuchs (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide* 36(1), 1–28.

Greenbaum, S. (1991). ICE: the International Corpus of English. *English Today* 7(4), 3.

Röthlisberger, M. The dative alternation in varieties of English. In preparation: KU Leuven PhD dissertation.

# LECTOMETRY AND LATENT VARIABLES

**Koen Plevoets**

**University of Leuven**

**Keywords:** ***latent variable models, correlation models, association models, register analysis, corpus linguistics***

Ever since its first formulation in Geeraerts, Grondelaers and Speelman (1999), lectometry has been widely used to map distances between language varieties or 'lects'. Often, these distances are given a geometrical representation in a low-dimensional space. Examples are the use of Multidimensional Scaling in Speelman, Grondelaers and Geeraerts (2003) and Ruette et al. (2014) and of Correspondence Analysis in Plevoets (2008), Delaere, De Sutter and Plevoets (2012), Prieels et al. (2015) and Ghyselen (2016). Usually, the number of dimensions of the geometrical space is chosen on the basis of representativeness, leading to an approximate picture of the linguistic variation. However, the spatial dimensions can also be interpreted as underlying factors governing the variability of the data. This methodological paper will explore this functional interpretation of the geometrical dimensions by establishing the link between lectometry and Latent Variable Models. It will be shown that the dimensions of the lectal space can be considered as hidden variables which lay bare specific causal mechanisms. In particular, analyses of translation and interpreting data will demonstrate that the lectometrical dimensions can be made to correspond to various socio-cultural determinants. That opens up the possibility for lectometrical studies of determining the 'social meaning' of linguistic varieties and variants.

**References:**

Delaere, I., G. De Sutter and K. Plevoets (2012). Is translated language more standardized than non-translated language? Using profile-based correspondence analysis for measuring linguistic distances between language varieties. *Target* 24 (2), 203–224.

Geeraerts, D., S. Grondelaers and D. Speelman (1999). *Convergentie en divergentie in de Nederlandse woordenschat: een onderzoek naar kleding- en voetbaltermen*. Amsterdam: P.J. Meertens-Instituut.

Ghyselen, A.-S. (2016). From diglossia to diaglossia: a West Flemish case-study. In M.-H. Côté, R. Knooihuizen and J. Nerbonne (eds.). *The Future of Dialects* (pp. 35–62). Berlin: Language Science Press.

Plevoets, K. (2008). *Tussen spreek- en standaardtaal. Een corpusgebaseerd onderzoek naar de situationele, regionale en sociale verspreiding van enkele morfosyntactische verschijnselen uit het gesproken Belgisch-Nederlands*. Leuven: Doctoral Dissertation.

Prieels, L., I. Delaere, K. Plevoets and G. De Sutter (2015). A corpus-based multivariate analysis of linguistic norm-adherence in audiovisual and written translation. *Across Languages and Cultures* 16(2), 209–231.

Ruette, T., D. Geeraerts, Y. Peirsman and D. Speelman (2014). Semantic weighting mechanisms in scalable lexical sociolectometry. In B. Szmrecsanyi and B. Wälchli (eds.). *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech* (pp. 205–230). Berlin/New York: De Gruyter.

Speelman, D., S. Grondelaers and D. Geeraerts (2003). Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities* 37(3), 317–337.

## CHARACTERIZING DIALECT GROUPS: CORRELATION AND INFORMATIVENESS ASSOCIATED WITH LINGUISTIC FORMS

**Gotzon Aurrekoetxea**

**Universidad del País Vasco**

**Esteve Clua**

**Universitat Pompeu Fabra**

**Aitor Iglesias**

**Universidad del País Vasco**

**Iker Usobiaga**

**Universidad del País Vasco**

**Miquel Salicrú**

**Universitat de Barcelona**

**Keywords:** ***dialectometry, MDS, representative and distinctive forms, central and border populations***

In dialectometry, attention is focused on identifying and characterizing dialects, interpreting spatial differences, and studying linguistic evolution over time. Based on a distance in which similarities and differences between populations are highlighted, fuzzy classification allows perimeters among dialect groups to be established and border/transition populations to be identified. The characterization of dialectal varieties requires processing a great deal of information. In this context, obtaining the reference populations of each group (central or pattern populations) and the most significant and different forms have allowed for focus to be turned to the most a priori relevant aspects. On a practical level, the direct application of this approach is questionable, because in dialectal corpuses which contain many redundant forms, analysis can be reduced to a set of (very informative and correlated) forms that explain only a part of the variation. The dependence on information provided by linguistic forms has been shown in multiple environments; for example, in some Romance languages affinity can be seen, among others, in certain verbal forms of the present indicative (second and third person singular and third person plural; first and second person plural,...).

In order to obtain a subset of forms which still maintains a significant percentage of the global information while presenting less redundancy, we carried the following steps: a) defining an affinity measure between forms, based on the correlation between interdistances ($d(F_i,F_j)=1-\rho^2(F_i,F_j)$); b) classifying forms and representing them on a 2D space (MDS with double label, group and form); c) choosing the most informative form of the most informative groups; and d) complementing the subset with equivalent forms (pertaining to the group) whose variation is likely to be governed by different rules from those of the previously selected most informative forms.

The Basque data used in this contribution has been taken from the "Recueil des idiomes de la Région Gasconne" compiled by Edouard Bourciez in 1895 (Aurrekoetxea and Videgain (2004)). The features of this corpus, structured as relational database, can be summarized as follows: 135 lexical concepts, 28 features of noun morphology, 24 about verb morphology, 23 about syntax and 26 diachronic features. This corpus has been analyzed as a linguistic atlas in Aurrekoetxea, Videgain and Iglesias (2004 and 2005) and in a dialectometric way in Aurrekoetxea and Videgain (2009), among others. Some clean-up processes have been performed on the data, and have been carried out in different ways: Firstly removing orthographical differences, secondly removing grammatical suffixes from the lexic, thirdly standardizing distinct word separations and, finally, repairing typographical errors.

**References:**


Aurrekoetxea, G. and X. Videgain (2004). Haur prodigoaren parabola Ipar Euskal Herriko 150 bertsiotan, Bilbao: UPV/EHU. Supplement of ASJU, XLIX.

Aurrekoetxea, G., A. Iglesias and X. Videgain (2004). Bourciez Bildumako Euskal Atlasa (BBEA-2): 1. Lexikoa. [Bourciez Linguistic Atlas: 1. Lexicon], ASJU 38-2 (2004) [ed. 2007].

Aurrekoetxea, G., A. Iglesias and X. Videgain (2005). Bourciez Bildumako Euskal Atlasa (BBEA-2): 2. Gramatika. [Bourciez Linguistic Atlas: 2. Grammar], ASJU 39-1 [ed. 2008].

Aurrekoetxea, G. and X. Videgain (2009). Le projet Bourciez: Traitement géolinguistique d'un corpus dialectal de 1895. *Dialectologia* 2, 81-111.

Bezdek, J. C. (2013). *Pattern recognition with fuzzy objective function algorithms*. New York: Springer Science & Business Media.

Clua, E. and M. Salicrú (2016a). Characterization of dialectal varieties: central and borders populations. Under review.

Clua, E. and M. Salicrú (2016b). New perspectives for analysis of dialect distance. *CILG* 2016.

Prokić, J., Ç. Çöltekin and J. Nerbonne (2012). Detecting Shibboleths. In M. Butt and J. Prokić (eds.) *Visualization of Language Patters and Uncovering Language History from Multilingual Resources*.+Workshop at the 13th Conference of the European Chapter of the Association for computational Linguistics. Avignon, France, 72-80.